KWALITEITSCENTRUM DIAGNOSTIEK VZW

Evaluatie van de Vlaamse normering van de WISC-V-NL

Kwaliteitscentrum voor Diagnostiek vzw
Kortrijksesteenweg 129
9000 Gent

Website: www.kwaliteitscentrumdiagnostiek.be
E-mail: communicatie@kwaliteitscentrumdiagnostiek.be

Titel:	Evaluatie van de Vlaamse normering van de WISC-V-NL
Redactie:	Kwaliteitscentrum voor Diagnostiek vzw
Datum:	Juni 2019

Dit rapport kwam tot stand met de steun van de Vlaamse Overheid. In deze tekst komen onderzoeksresultaten van de auteur(s) naar voor en niet van de Vlaamse Overheid. Het Vlaams Gewest kan niet aansprakelijk gesteld worden voor het gebruik dat kan worden gemaakt van de meegedeelde gegevens. Niets uit deze uitgave mag worden verveelvoudigd en/of openbaar gemaakt zonder uitdrukkelijk te verwijzen naar de bron.

Reviewprocedure

De evaluatie van de Vlaamse normering van de WISC-V-NL kwam tot stand na raadpleging van twee externe, onafhankelijke beoordelaars met expertise inzake testconstructie, psychometrie en het meten van cognitieve vaardigheden. Beide externe beoordelaars hebben het testmateriaal en de handleidingen van de WISC-V-NL grondig onderzocht en een gemotiveerde beoordeling ervan bezorgd aan het Kwaliteitscentrum voor Diagnostiek vzw. Daarnaast heeft een onderzoeker van het Kwaliteitscentrum voor Diagnostiek vzw onafhankelijk een beoordeling gemaakt op basis van een analyse van de Technische handleiding. Op basis van deze drie beoordelingen werd de testauteur Pearson Benelux B.V. gecontacteerd met enkele vragen voor bijkomende informatie, waarna vervolgens door het Kwaliteitscentrum voor Diagnostiek vzw een eerste versie van het reviewrapport opgemaakt werd. De drie onafhankelijke beoordelingen en de bijkomende informatie aangeleverd door de testauteur vormden hiervoor de basis. Deze eerste versie werd in een volgende stap ter discussie voorgelegd aan de externe beoordelaars, waarna nog verschillende aanpassingen gebeurden en een aangepaste versie van het reviewrapport gefinaliseerd werd. Vervolgens werd deze aangepaste versie van het reviewrapport ter inzage voorgelegd aan de testauteur. Via een gemotiveerd schrijven heeft de testauteur gereageerd op de inhoud van het reviewrapport en extra toelichting gegeven omtrent enkele opmerkingen. Op basis van deze inhoudelijke toelichting door de testauteur werd in onderling overleg tussen de beoordelaars besloten om enkele zaken uit het reviewrapport aan te passen. Dit resulteerde vervolgens in de finale versie van het reviewrapport.

Als kader bij de beoordeling dienden de richtlijnen inzake normering, validiteit en betrouwbaarheid zoals geformuleerd door de Commissie Testaangelegenheden Nederland (COTAN; Evers, Lucassen, Meijer, & Sijtsma, 2010), de European Federation of Psychologists' Associations (EFPA; Evers et al., 2013) en de Sectie Psychodiagnostiek van de Belgische Federatie van Psychologen (BFP).

Dit rapport heeft betrekking op het Vlaamse normeringsonderzoek van de WISC-V-NL en handelt niet over de Nederlandse data. De beoordeling van de Nederlandse normering van de WISC-V-NL wordt verricht door de Commissie Testaangelegenheden Nederland (COTAN). Voor het bekomen van het adviesrapport van de COTAN verwijzen we door naar: https://www.cotandocumentatie.nl/beoordelingen/b/15651/wechsler-intelligence-scale-for-children-fifth-edition/. In Vlaanderen wordt gebruikgemaakt van de gecombineerde Nederlands-Vlaamse normen. Daarom dient de beoordeling van de COTAN ook in overweging genomen te worden om te komen tot een verantwoord gebruik van de WISC-V-NL in Vlaanderen.

Algemeen

Jaar van uitgave:	2018
Auteur(s):	D. Wechsler
Bewerker(s):	M.P.H. Hendriks, S. Ruiter, M. Schittekatte, A. Bos
Uitgever:	Pearson Assessment and Information B.V.
Referenties:	Hendriks, M.P.H., Ruiter, S., Schittekatte, M., & Bos, A. (2018). *Wechsler Intelligence Scale for Children – Fifth edition – Nederlandstalige bewerking. Technische handleiding.* Amsterdam: Pearson Benelux B.V. Wechsler, D. (2017). *Wechsler Intelligence Scale for Children – Fifth edition – Nederlandstalige bewerking. Afname- en scoringshandleiding.* Amsterdam: Pearson Benelux B.V.

Meetpretentie en doelgroep

De Wechsler Intelligence Scale for Children, 5de editie - Nederlandstalige bewerking (WISC-V-NL) is een instrument dat de cognitieve vaardigheden in kaart brengt van kinderen en jongeren van 6 tot en met 16 jaar.

Structuur

De Nederlandstalige WISC-V bestaat uit 14 subtests (tien primaire en vier secundaire subtests), namelijk *Blokpatronen*, *Overeenkomsten*, *Matrix Redeneren*, *Cijferreeksen*, *Symbool Substitutie Coderen*, *Woordenschat*, *Gewichten*, *Figuur samenstellen*, *Plaatjesreeksen*, *Symbool Zoeken*, *Cijfers en Letters Nazeggen*, *Figuur Zoeken*, *Begrijpen* en *Rekenen*. De tien primaire subtests dragen bij tot vijf primaire indexscores: Verbaal Begrip Index, Visueel Ruimtelijke Index, Fluïde Redeneren Index, Werkgeheugen Index en Verwerkingssnelheid Index. Het Totaal IQ kan op basis van zeven (van de tien) primaire subtests berekend worden. Daarnaast zijn er vijf aanvullende indexen die gebaseerd worden op combinaties van primaire of primaire en secundaire subtests. De aanvullende indexen zijn: Kwantitatief Redeneren Index, Auditief Werkgeheugen Index, Non-Verbale Index, Algemene Vaardigheid Index en Cognitieve Competentie Index. Als laatste kunnen ook geschaalde processcores berekend worden bij de subtests Blokpatronen, Cijferreeksen en Figuur Zoeken. Deze processcores dragen niet bij aan het Totaal IQ, maar kunnen een dieper inzicht verschaffen over de cognitieve processen die meegespeeld hebben tijdens de prestatie op de subtest.

Afname en scoring

De afnameduur van de WISC-V-NL is gemiddeld twee uur, afhankelijk van de leeftijd en capaciteiten van het kind of de jongere.

Er zijn twee manieren om de WISC-V-NL af te nemen. Naast een papier-en-potlood versie bestaat er ook een digitale versie (Q-interactive). Het gaat om een nieuwe afnamewijze waarbij proefleider

en kind elk over een tablet beschikken, tegenover elkaar zitten en via de tablets communiceren. Al het testmateriaal is beschikbaar via het scherm met uitzondering van de subtests *Blokpatronen* en *Figuur Zoeken*. Bij de subtest *Blokpatronen* wordt er gebruik gemaakt van fysieke blokken. Het tonen van de stimuli, tijdswaarneming en scoring gebeuren wel digitaal. De subtest *Figuur Zoeken* wordt nog op papier afgenomen, maar de tijdswaarneming en scoring daarvan kan wel plaatsvinden via Q-interactive.

De scoring kan manueel (papier-en-potlood) aan de hand van de voorziene scoringsformulieren. Als alternatief biedt de testauteur een betalend online platform (Q-Global) aan waarbij het volstaat de bekomen ruwe scores in te brengen. Het platform genereert automatisch een omstandig rapport met alle normscores, base-rate, significanties en dergelijke meer. Dit vergemakkelijkt aanzienlijk het administratieve en opzoekingswerk. Ook bij het gebruik van Q-interactive gebeurt de scoring automatisch. In het kader van de beoordeling was het niet mogelijk om de juistheid van de software te controleren. De testauteur heeft hierover desgevraagd aangegeven dat de resultaten zoals die gegenereerd kunnen worden met behulp van Q-global, volledig overeenkomstig zijn met manuele scoring met behulp van de Afname- en scoringshandleiding. Dezelfde tabellen als in de handleiding zijn opgenomen en ingebouwd in Q-global. Er wordt middels een geautomatiseerd kwaliteitsproces een database van duizenden cases gebruikt om de resultaten te vergelijken en te controleren. Daarnaast zijn er nog eens handmatig tientallen cases getest om te verzekeren dat de resultaten exact overeenkomen.

Doorheen de testverwerking worden in hoofdzaak vier referentie metrieken gehanteerd:

- genormaliseerde standaardscores (geschaalde scores) met gemiddelde = 10 en standaarddeviatie = 3 voor de primaire en secundaire subtests en geschaalde processcores;
- genormaliseerde standaardscores met gemiddelde = 100 en standaarddeviatie = 15 bij het Totaal IQ en de Indexscores;
- base-rates o.a. bij de ruwe scores bij de processcores;
- ontwikkelingsleeftijdsequivalenten van de ruwe subtestscores.

Het vraagt van de diagnosticus een gedegen inzicht in de specifieke eigenschappen van deze scores om er op een gepaste wijze mee te kunnen omgaan bij de interpretatie.

Afname

Afname:	Individueel
Wijze:	Papier & potlood, Q-interactive
Door:	Psychologisch werker, Psychodiagnostisch medewerker, Toegepast (hbo-)psycholoog en Psychologisch assistent
Tijdsduur:	120 min

Scoring

Papier en potlood:	Handmatig
Q-global:	Automatisch
Q-interactive:	Automatisch

Interpretatie

Door:	Diagnostisch gekwalificeerde (ortho)pedagoog en psycholoog

Het is belangrijk op te merken dat er enkel Vlaamse en Nederlandse normeringsdata verzameld zijn aan de hand van de papier-en-potlood versie. Dit rapport heeft dan ook enkel betrekking op de kwaliteit van de Vlaamse normen aan de hand van de papier-en-potlood versie en is dus niet van toepassing op de digitale afname Q-interactive in Vlaanderen.

Beoordeling

Uitgangspunten testconstructie

Met de huidige WISC-V wordt de jarenlange vertrouwde opsplitsing in Verbaal en Performaal IQ verlaten. Er wordt een hiërarchisch model voorgesteld met drie interpretatieniveaus: totaalscore, primaire indexen en aanvullende indexen.

Al refereert de Technische handleiding eerder weinig expliciet naar het CHC-model (Cattell-Horn-Carroll) is het duidelijk dat de test dit model operationaliseert (Flanagan & Alfonso, 2017; Kaufman, Riaford, & Coalson, 2016; Weis, Saklofske, Holdnack, & Prifitera, 2016). Onderstaande vergelijking met betrekking tot de primaire indexen maakt dit duidelijk.

WISC-V Primaire Index		**CHC Brede Cognitieve Vaardigheid Index**
Verbale Begrip Index (VBI)	→	Gekristalliseerde intelligentie (Gc)
Visueel Ruimtelijke Index (VRI)	→	Visuele Informatieverwerking (Gv)
Fluid Redeneren Index (FRI)	→	Vloeiende intelligentie (Gf)
Werkgeheugen Index (Wgl)	→	Kortetermijngeheugen (Gsm)
Verwerkingssnelheid Index (VsI)	→	Verwerkingssnelheid (Gs)

De handleiding vermeldt nergens waarom bij het totaal IQ gekozen wordt (in CHC-termen) voor twee Gc subtests, twee Gf subtests, één Gsm subtest, één Gv subtest en één Gs subtest, al is dit gemakkelijk af te leiden uit het CHC-model (zie o.a. ook CoVaT-CHC). Binnen het CHC-model wordt eveneens gesteld dat de Brede Cognitieve Vaardigheid (BCV)-Indexen (tweede 'stratum') zich best steunen op minstens twee subtests, liefst elk van een verschillende onderliggende Nauwe Cognitieve Vaardigheid (NCV). De WISC-V-NL doet dit mooi, maar vermeldt dus niet het waarom. Vanuit het CHC-gedachtegoed kan voorgesteld worden om standaard de tien subtests toe te passen zodat men naast het IQ (op de zeven subtests) ook de vijf BCV-indexen in kaart brengt.

Naast het IQ en de vijf primaire indexen verschaft de test nog vijf aanvullende indexen. Met deze indexen verlaat de WISC-V het CHC-referentiekader en men merkt direct het eerdere speculatieve inhoudelijke karakter ervan. Het is duidelijk minder onderbouwd. Bijvoorbeeld:

- De Kwantitatief Redeneren Index (KRI) omvat *Gewichten* en *Rekenen* terwijl er momenteel onvoldoende evidentie is dat beide als één construct gezien kunnen worden.
- De Non Verbale Index (NVI) omvat subtests die geen taal vergen in de uitvoering, maar heeft wel talige instructies. Niet-talige instructies, zoals bij de Snijders-Oomen Niet-Verbale Intelligentietests (SON-R 6-40, SON-R 2.5-7, SON-R 5.5-17), zijn niet gestandaardiseerd aanwezig.
- De Algemene Vaardigheid Index (AVI) presenteert zich met de subtests (*Overeenkomsten*, *Woordenschat*, *Blokpatronen*, *Matrix Redeneren* en *Gewichten*) vooral als een – in CHC-termen – Gf/Gc Index. De correlatie met het IQ kan niet anders dan hoog zijn (vijf subtests van de zeven). De ervaring zal moeten aantonen hoe zinnig deze Index is binnen het hele plaatje en waarom men Gsm en Gs niet betrokken wenst te zien. De verleiding is groot om de AVI als een tweede IQ te zien, wat het niet is.

Naast dit alles introduceert de WISC-V de Processcores (ook aanwezig in de WAIS-IV). Deze benadering draagt duidelijk de stempel van de neuropsychologische invalshoek. De WISC-V mikt

veel meer dan de vorige versies op toepasbaarheid binnen dit onderzoeksdomein. Onrechtstreeks introduceert de test ook het CHC-model in het neuropsychologisch denken, een aanzienlijke verdienste. De Processcores geven aanvullende gekwantificeerde informatie binnen bepaalde subtests zoals voor- versus achterwaarts herhalen bij de subtest *Cijferreeksen*.

Algemeen genomen kan men de **uitgangspunten van de testconstructie** als **goed** beoordelen.

Kwaliteit van het testmateriaal

Papier-en-potloodversie

De testkoffer omvat het concrete testmateriaal, de scoringsformulieren en twee handleidingen (nl. de Afname- en scoringshandleiding en de Technische handleiding).

Het testmateriaal is hedendaags, (voldoende) aansprekend en verzorgd. Belangrijk hierbij is de mate van interculturele fairness. In de USA is het testmateriaal uitvoerig hierop getest. Bij de Nederlandstalige versie werd hieraan, o.a. in samenwerking met experten rond dit thema, aanvullend bijzondere aandacht besteed. We kunnen stellen dat er voldoende inspanningen geleverd zijn om de interculturele fairheid van het instrument te maximaliseren en de bias – ook op item niveau – te minimaliseren. Dit betekent echter niet dat de test culture-free is, ook niet de Non Verbale Index. Er bestaan gewoonweg geen culture-free (intelligentie)tests, wél tests die hieraan aandacht besteden. De WISC-V is daar één van en doet dit behoorlijk. Toch is extra onderzoek noodzakelijk om echte conclusies te kunnen maken omtrent fairness en mogelijke bias. Differential item functioning (DIF) op basis van de normdata zou bijvoorbeeld mogelijk kunnen zijn.

Het is jammer dat er in de Technische handleiding geen gegevens beschikbaar zijn over de moeilijkheidsgraden van de items. Deze zijn belangrijk om te kunnen beoordelen of de afbreek- en omkeerregels juist functioneren.

Computerversie (Q-interactive)

De computerversie werd niet beoordeeld aangezien de Nederlandse en Vlaamse normering gebeurd is aan de hand van de papier-en-potloodversie.

Aandachtspunten

- Ondanks de aandacht die besteed is aan de interculturele fairness van de test, is binnen het Nederlands-Vlaams normeringsonderzoek onvoldoende aangetoond dat alle gehanteerde items niet cultuur-geladen zijn.
- Er is geen Nederlands-Vlaams onderzoek verricht met de digitale afname van de WISC-V-NL. De huidige normen zijn gebaseerd op de papier-en-potlood versie van de test.

Algemeen genomen kan men de **kwaliteit van het testmateriaal** als **goed** beoordelen. Bovenstaande aandachtspunten zijn evenwel van toepassing.

Kwaliteit van de handleiding

De beide handleidingen hebben een uitstekende kwaliteit. Voor Vlaanderen is de overgang van de WISC-III-NL naar de WISC-V-NL een grote sprong, des te meer omdat de WISC-IV, die dit als het ware voorbereidde, nooit voor Vlaanderen en Nederland beschikbaar was. De Vlaamse gebruiker zal wel vertrouwde subtests tegenkomen, maar er zijn ook nieuwe subtests, en vertrouwde subtests hebben soms nieuwe afname- en scoringsinstructies. De Afname- en scoringshandleiding is dan ook onmisbaar voor een gestandaardiseerde afname en correcte scoring. Daarnaast is de Technische handleiding tevens noodzakelijk om zich echt vertrouwd te maken met de test en met de interpretatie van de scores. Het doornemen van deze handleiding is voor alle practici een must.

Zoals blijkt uit de bespreking doorheen dit reviewrapport werden enkele onvolledigheden opgemerkt in beide handleidingen (cf. infra en supra). Het ontbreken van volgende informatie willen we extra onder de aandacht brengen:

- gegevens over de moeilijkheidsgraden van de items in de Technische handleiding;
- de basistabel met de gemiddelde ruwe subtestscores en hun standaarddeviatie in de Technische handleiding;
- het benadrukken van de verwarring tussen 120 seconden en 1 minuut 20 seconden in de Afname- en scoringshandleiding gezien dit bij het Vlaamse normeringsonderzoek tot een foutieve afname van de subtests *Symbool Substitutie Coderen* en *Symbool Zoeken* heeft geleid.

Algemeen genomen kan men de kwaliteit van de handleiding als **goed** beoordelen.

Vlaamse normen

Het normeringsonderzoek is gebaseerd op 1433 kinderen en jongeren, 1038 uit Nederland en 395 uit Vlaanderen. Er zijn zowel Nederlandse, als Nederlands-Vlaamse normen beschikbaar. Opvallend hierbij is dat wegens een systematische testleider inschattingsfout in de loop van de normering van de subtests *Symbool Zoeken* en *Symbool Substitutie Coderen* enkel Nederlandse normen beschikbaar zijn voor deze subtests, dus ook binnen de Nederlands-Vlaamse normtabellen (cf. infra).

In wat volgt wordt apart ingegaan op de Vlaamse normeringssteekproef. Het kwaliteitsoordeel over de Vlaamse normering gaat ervan uit dat de Vlaamse data samengevoegd worden met de Nederlandse data. De Vlaamse normeringssteekproef is te klein om aparte kwaliteitsvolle normen te kunnen genereren.

Algemeen

De Vlaamse steekproef bestond uit 395 kinderen en jongeren, gestratificeerd volgens leeftijd, geslacht, opleidingsniveau moeder, opleidingsniveau kind, onderwijsnet, nationaliteit, regio en urbanisatiegraad. Van deze stratificatievariabelen werden populatiecijfers aangeleverd door het Kwaliteitscentrum voor Diagnostiek vzw[1]. Er werd tijdens de dataverzameling getracht deze streefpercentages zo goed mogelijk te benaderen. Aansluitend heeft er bij de data-analyse een weging plaatsgevonden, om de steekproef waar nodig meer in overeenstemming te brengen met

[1] De populatiecijfers beschreven in de handleiding komen niet exact overeenkomen met de aangeleverde cijfers door het Kwaliteitscentrum voor Diagnostiek vzw. Er zijn minieme verschillen waar te nemen.

de populatie. De gehanteerde wegingsfactoren waren niet groter dan 2, zoals vereist in de COTAN-richtlijnen (Evers et al., 2010). Na een optimalisatie en weging werd een steekproefgrootte van 361 bereikt. Dit stemt overeen met het door het Kwaliteitscentrum voor Diagnostiek vzw vereiste aantal van 356 personen. Het is belangrijk op te merken dat deze vooropgestelde steekproefgrootte gebaseerd is op de veronderstelling dat de Vlaamse steekproef zou worden samengevoegd met de Nederlandse steekproef indien de verschillen tussen Vlaanderen en Nederland minimaal zouden zijn.

Dataverzameling

De dataverzameling in Vlaanderen is gebeurd in de periode oktober 2016 – juni 2017. De data werden hoofdzakelijk verzameld door studenten van drie verschillende onderwijsinstellingen. Positief hierbij is de grote zorg die besteed werd aan de gestandaardiseerde opleiding van elke student. Daarnaast werden op het einde van de dataverzameling nog een veertigtal afnames gerealiseerd door ervaren testleiders. Deze ervaren testleiders waren enkele docenten en assistenten die instonden voor de opleiding van de studenten. Ook de data van de klinische groepen werden verzameld door ervaren testleiders. Het is mogelijk dat er bij de dataverzameling sprake was van een systematisch testleider-effect, enerzijds op het niveau van de onderwijsinstelling, anderzijds op het niveau van ervaring van de testleider. Vooral het verschil in ervaring van de testleider baart ons enige zorgen. Zeker gezien het feit dat de afnames door ervaren testleiders gerealiseerd werden bij klinische groepen en bij jongeren met laagopgeleide moeders kan een mogelijke testleiderbias een systematisch effect in de data creëren. Deze hypothese kon op basis van de dataset niet verder onderzocht en uitgesloten worden.

Representativiteit

In de Technische handleiding wordt per variabele de vergelijking gemaakt tussen het streefpercentage (populatie) en het gewogen steekproefpercentage (Tabellen 4.11b, 4.12b, 4.13b, 4.14b, 4.15b). De afwijkingen van de streefpercentages worden acceptabel genoemd. Het is echter onduidelijk welke definitie voor 'acceptabel' wordt gehanteerd. Bijvoorbeeld, bij de ontwikkeling van normen voor de Bayley-III-NL werden alle percentages die zich situeren binnen het interval]streefpercentage - 5%, streefpercentage + 5%[beoordeeld als percentages die het streefpercentage 'voldoende' benaderen (van Baar, Steenis, Verhoeven, Hessen, & Smits-Engelsman, 2015). Het lijkt alsof bij de WISC-V-NL hetzelfde criterium wordt gehanteerd omdat af en toe een verwijzing naar een 5%-afwijking wordt gemaakt. Meer transparantie hierover is gewenst. Indien we dit criterium hanteren bij alle variabelen, zijn er (na weging) zeer lichte afwijkingen te observeren bij onderwijsnet (VGO: steekproefpercentage 68.8%, populatiepercentage 74.2%) en urbanisatie (stad: steekproefpercentage 50.1%, populatiepercentage 56.0%). Ook binnen de leeftijdsgroepen 12-jarigen en 16-jarigen is er een verschil van meer dan 5% tussen jongens en meisjes (12-jarigen: streefpercentage jongens 50%, populatiepercentage 57.7%; 16-jarigen: streefpercentage jongens 50%, populatiepercentage 42.5%).

Verder is het op basis van de tabellen (4.12b, 4.13b, 4.14b, 4.15b) niet mogelijk na te gaan of er in een bepaalde leeftijdscategorie een onder- of oververtegenwoordiging van kinderen met specifieke (stratificatie)kenmerken heeft plaatsgevonden. De verdeling volgens de stratificatiecriteria wordt namelijk nooit per leeftijdsgroep apart weergegeven (met als uitzondering de variabele geslacht). Op p. 63 §2 van de handleiding wordt enkel gesteld dat "de overige variabelen representatief gestratificeerd zijn per leeftijdsgroep". Meer informatie hieromtrent werd opgevraagd bij Pearson Benelux B.V. De testauteur leverde ons tabellen aan met de verdeling (na weging) per leeftijdsgroep voor de variabelen opleidingsniveau moeder,

nationaliteit en regio[2]. Opvallend aan deze data was dat er veel cellen waren waarbij de afwijking van het streefpercentage groter dan 5% was (24 op 99 cellen). Er zijn in de wetenschappelijke literatuur nog geen duidelijke richtlijnen over de representativiteit per leeftijdsgroep bij continue normering. Enerzijds wordt gesteld dat het belangrijk is dat de steekproefpercentages ook binnen elke leeftijdsgroep overeenkomen met de streefpercentages (Becher, Hemker, & Maris, 2009, p. 6). Anderzijds wordt bij continue normering vooral het belang van de uiterste leeftijdsgroepen onderstreept. De uiterste leeftijdsgroepen dienen groter te zijn dan de middelste groepen (Evers, Sijtsma, Lucassen, & Meijer, 2010). Het is dan zeker bij deze uiterste datapunten belangrijk dat er een goede vertegenwoordiging is van de populatie. Toegepast op de Vlaamse data kan gesteld worden dat de grotere dataverzameling bij de uiterste leeftijdsgroepen (en dan vooral bij de oudste leeftijdsgroep) niet gerealiseerd werd. Daarnaast ziet men in de extra tabellen door de testauteur aangeleverd dat er voor de oudste leeftijdsgroep (16:0-16:11) een ondervertegenwoordiging is van laag opleidingsniveau moeder en een oververtegenwoordiging van midden opleidingsniveau moeder en autochtoon. Er is dan ook voornamelijk bezorgdheid om de juiste stratificatie van deze uiterste groep. De normen van deze uiterste leeftijdsgroep moeten dan ook met de nodige voorzichtigheid gehanteerd worden.

Nederland versus Vlaanderen

In het oorspronkelijke steekproefplan is uitgegaan van een gewenste Vlaamse steekproefgrootte van 356 kinderen en jongeren, op voorwaarde dat er geen verschillen tussen Nederland en Vlaanderen geobserveerd worden en bijgevolg de Nederlandse en Vlaamse data kunnen samengevoegd worden voor de berekening van de normen. Om mogelijke verschillen tussen beide landen in ruwe testscores te onderzoeken werd beroep gedaan op enkele MANOVA's[3], zoals beschreven in de Technische handleiding. Er werd aan de testauteur gevraagd om bijkomende analyses op de data uit te voeren om het verschil tussen Nederland en Vlaanderen verder te onderzoeken. Meer bepaald werd gevraagd ook op subtestniveau en per leeftijdscategorie de verschillen tussen beide landen te onderzoeken. Voor de inflatie van type-1 fouten werd gecontroleerd. Afhankelijk van welke analyse gehanteerd werd, kwamen verschillende significante effecten aan het licht. Zonder in te gaan op de details van de analyses, worden hier de verschillende gevonden effecten op basis van de verschillende analyses opgesomd:

- een significant hoofdeffect van land op Totaal IQ;
- een significant hoofdeffect van land op de Werkgeheugen Index;
- een significant hoofdeffect van land en een significante interactie tussen land en leeftijd op de subtest *Cijferreeksen*;
- een significant hoofdeffect van land op de subtest *Plaatjesreeksen*;
- een significant hoofdeffect van land op de subtest *Cijfers en Letters Nazeggen*;
- een significant verschil tussen Nederland en Vlaanderen in de leeftijdsgroepen 9-10-jarigen en 11-12-jarigen bij de Werkgeheugen Index.

De effectgroottes van deze effecten zijn echter klein. Er kan dan ook gesteld worden dat de effecten zodanig klein zijn dat ze weinig tot geen invloed uitoefenen op de interpretatie van de resultaten. Zo kan bijvoorbeeld beargumenteerd worden dat bij een klein effect het verschil ondervangen kan

[2] Er werd geen informatie aangeleverd om te besluiten dat per leerjaar van het secundair onderwijs per opleidingsniveau van het kind de streefpercentages ook gehaald werden.
[3] Het is wenselijk om bij de gerapporteerde F-toetsen in de Technische handleiding ook de vrijheidsgraden te rapporteren.

worden door steeds gebruik te maken van betrouwbaarheidsintervallen bij de interpretatie van de testresultaten.

Symbool Substitutie Coderen en Symbool Zoeken

Uit de data-analyse is gebleken dat bij de subtests *Symbool Substitutie Coderen* en *Symbool Zoeken* een afwijkende afnamemethodiek gevolgd is geweest door bepaalde Vlaamse testleiders. In plaats van de test af te breken op 120 seconden hebben sommige testleiders de test afgebroken op 80 seconden (1 minuut 20 seconden). Om vast te stellen of er geen verschil is tussen Nederland en Vlaanderen wordt een extra analyse in de Technische handleiding beschreven waarin de data op 60 seconden vergeleken werd tussen beide landen. Deze analyse toonde geen effect van land voor beide subtests. Helaas kunnen er op basis van de huidige dataset geen extra analyses uitgevoerd worden om met meer zekerheid te stellen dat de Nederlandse en Vlaamse data niet verschillen van elkaar. Ook is het onduidelijk hoeveel testleiders van de Vlaamse normgroep de afwijkende afnamemethodiek hebben gehanteerd. Dit kan op basis van de huidige dataset niet exact bepaald worden.

Men moet zich er bij de interpretatie van de resultaten van deze subtests van bewust zijn dat enkel vergeleken kan worden met de Nederlandse normgroep en dit mogelijks een kleine afwijking inhoudt ten opzichte van de Vlaamse situatie. Verder willen we het belang van het correct en gestandaardiseerd afnemen van de subtests onder de aandacht brengen. Het lijkt erop dat het misverstand (tussen 120 seconden en 1 minuut 20 seconden) niet expliciet in de Afname- en scoringshandleiding werd aangekaart opdat in de toekomst dergelijke fouten minder zouden kunnen voorkomen.

Continue normering

Steekproefgrootte

De normering steunt op een continu normeringsmodel. In de handleiding wordt een vergelijking beschreven tussen klassieke en continue normering volgens de methode van Bechger et al., (2009). Deze methode werd in het verleden bij testconstructie (o.a., Bayley-III-NL) reeds aangehaald om aan te tonen dat het gebruikte aantal proefpersonen binnen elke leeftijdsgroep in het continue normeringsmodel voldoende is.

Aan de hand van de methode van Bechger et al. (2009) werd de standaardfout van zeven geschaalde subtestscores (die bijdragen aan het Totaal IQ) per leeftijdsgroep berekend voor twee klassieke normeringsmodellen (N = 300 en N = 400), en vergeleken met de standaardfout onder het continu normeringsmodel. Dit werd zowel gedaan voor de Nederlandse data, als de gecombineerde Nederlands-Vlaamse data. Op basis van deze vergelijkingen kon aangetoond worden dat voor beide datasets het continu normeringsmodel voor elke leeftijdsgroep kleinere schattingen van de standaardfout van de geschaalde subtestscores opleverde dan een klassiek normeringsmodel met N = 300. Op basis van deze resultaten wordt door de testauteur besloten dat beide steekproefgroottes van voldoende omvang zijn. Belangrijk om te vermelden is dat de Vlaamse steekproef op zich te weinig proefpersonen bevat om aparte Vlaamse normen te ontwikkelen.

Ook binnen de COTAN-richtlijnen wordt de methode van Bechger et al. (2009) voor de vergelijking tussen klassieke en continue normering beschreven (Evers et al., 2010). Evers et al. (2010) wijst er echter op dat bij het hanteren van de methode van Bechger et al. er rekening moet gehouden worden met een aantal statistische veronderstellingen (e.g., homoscedasticiteit, normaliteit). Er dient dan ook in de handleiding van de WISC-V-NL eerst aangetoond te worden dat de huidige

dataset voldoet aan deze veronderstellingen. Verder werden beide normeringsmethoden enkel met elkaar vergeleken op basis van de standaardfout van het gemiddelde. Dit dient ook nog verder onderzocht te worden voor andere parameters van de verdeling, zoals de standaarddeviatie en scheefheid van de verdeling (Evers et al., 2010).

Virtueel karakter extreem hoge en lage IQ-scores

De normen werden ontwikkeld met een inferentieel continu normeringsmodel. Dit model wordt omschreven als: "*naast de gelijkenis met continu normeren verschilt inferentieel normeren hiervan omdat gebruik gemaakt wordt van de Johnsoncurve om de theoretische verdeling af te leiden. Op deze manier kunnen de frequenties bepaald worden van extreme scores (extreem lage of extreme hoge scores) die niet altijd geobserveerd worden in een steekproef*" (Technische handleiding p. 66).

Het gaat hier om een aanslepende discussie die al stamt uit de eerste Stanford-Binet (1916). Men wil ook graag zeer hoge IQ's en zeer lage IQ's bepalen. In feite "werken" echter onze moderne IQ-tests enkel tussen 50/55 tot en met 145/150. Daarboven en daaronder worden IQ's virtueel van aard. De tests zijn weinig tot niet geschikt om de zeer lage (-50) en de zeer hoge (150+) IQ's effectief te meten. Andere instrumenten of metrische aanpak zijn dan nodig. Statistiek laat echter wel toe om een soort extrapolatie te maken, maar dat is nagenoeg een meetmathematische speculatie. Zo heeft bijvoorbeeld een IQ van 159 en hoger een aanwezigheidsfrequentie van 4 personen op 100000 (0.0042%). Als men weet dat de representatieve normering voor de WISC-V-NL rond de 1500 personen ligt, ziet men dat de kans quasi 0 is om één persoon met een IQ van 160 (of IQ van 40) in de normgroep te hebben. De uitspraak: "*de Johnsoncurve laat toe zeer extreem hoge/lage scores, die niet altijd geobserveerd worden in een steekproef, te bepalen*", is relevant (voor het virtuele karakter daarvan). De problematiek is niet specifiek voor de WISC-V alleen. Wel specifiek is dat bij de WISC-V de IQ-tabellen gaan tot respectievelijk IQ 40 en IQ 160 en zo de indruk gegeven wordt dat de test effectief metrisch bruikbaar is in de extreme zone. Dat het anders kan bewijzen de WPPSI-III en WAIS-IV waarbij de IQ-tabellen niet lager gaan dan IQ 55 en niet hoger dan IQ 145.

Een knelpunt dat bij dit alles onmiddellijk aansluit is de paragraaf op p. 143 uit de Technische handleiding: "*in Tabel 8.3 wordt zichtbaar dat er in de normeringssteekproef iets te weinig kinderen in het gebied zeer hoog vallen. Dit gaat om slechts 1.2% te weinig, wat betekent dat 15 kinderen (1.2% maal 1291) meer hadden moeten getest worden*." Ten eerste is die 1291 onduidelijk, elders kan men afleiden dat de normgroep 1396 personen (gewogen; N = 1433 ongewogen) omvatte, wat dan 17 kinderen betekent. Ten tweede gaat het erom dat er 15 hoog begaafde kinderen moeten bijkomen. Dit betekent dat bij een representatieve/toevallige steekproef er wel heel wat meer kinderen zouden bij betrokken moeten worden om die 15 eruit te distilleren. In feite ontstaat echter zo een cirkelredenering waarbij ten opzichte van de algemene populatie het tekort structureel blijft en de vraag naar de optimaliteit van het meten in die zone door de WISC-V naar voren komt. Dat dit volgens de handleiding weinig effect zou hebben op de normen lijkt verdedigbaar, maar het ondersteunt het feitelijk virtuele karakter van de WISC-V in de extreme IQ-zone.

Aandachtspunten

- Er werd bij het Vlaamse normeringsonderzoek grote zorg besteed aan een correct gestratificeerde dataverzameling. Toch kunnen na het analyseren van de data enkele tekorten opgemerkt worden. Enerzijds werd de aanbeveling om meer data te verzamelen in de uiterste leeftijdsgroepen niet helemaal gerealiseerd. Anderzijds vielen ook tekorten op wat betreft de stratificatie (volgens opleidingsniveau moeder, nationaliteit en regio) per leeftijdsgroep. Voornamelijk de mindere overeenkomst tussen de populatie- en steekproefcijfers wat betreft opleidingsniveau moeder en nationaliteit in de oudste leeftijdsgroep (16:0-16:11) is hierbij van belang. Voorzichtigheid is geboden bij de interpretatie van de normen in deze uiterste leeftijdsgroep.
- Verschillende analyses tonen enkele significante, doch kleine verschillen tussen Nederland en Vlaanderen op bepaalde subtest/index/totaalscores. Gezien de kleine effectgrootte van deze verschillen kan beargumenteerd worden dat de Nederlandse en Vlaamse data samengenomen mogen worden bij de constructie van de normen. Het gebruik van betrouwbaarheidsintervallen bij de interpretatie van de scores kan de minieme verschillen tussen Nederland en Vlaanderen ondervangen.
- Hanteer de correcte tijdslimiet van 120 seconden bij de subtesten *Symbool Substitutie Coderen* en *Symbool Zoeken*. Verwarring kan ontstaan tussen de instructie van 120 seconden (i.e., 2 minuten) en de notatie van 1 minuut 20 seconden bij de meeste stopwatches. Bijkomend moet men er zich bij de interpretatie van de resultaten van deze subtesten van bewust zijn dat enkel Nederlandse normdata voorhanden is. Er waren geen aanvullende analyses mogelijk om met meer zekerheid de equivalentie tussen de Nederlandse en Vlaamse normdata te garanderen.
- Ook al gaan de normtabellen van de WISC-V-NL tot respectievelijk IQ 40 en IQ 160, het instrument is weinig tot niet geschikt om de zeer lage (-50) en de zeer hoge (150+) IQ's effectief te meten. Bij zeer hoge en lage IQ's wordt aanbevolen meer aangepaste instrumenten en/of methodieken te gebruiken.

Algemeen genomen kan men de **Vlaamse normering** als **voldoende** beoordelen. Bovenstaande aandachtspunten zijn evenwel van toepassing. De Nederlandse normeringssteekproef werd beoordeeld door de COTAN. Na consultatie van deze COTAN-beoordeling zijn we van oordeel dat het kwaliteitslabel voor de **gecombineerde Nederlands-Vlaamse normen** gelijklopend is aan dit van de Vlaamse normering. Voor details van de COTAN-beoordeling verwijzen we door naar: https://www.cotandocumentatie.nl/beoordelingen/b/15651/wechsler-intelligence-scale-for-children-fifth-edition/.

Betrouwbaarheid

Wat de beoordeling van de betrouwbaarheidscoëfficiënten betreft, worden verschillende criteria gehanteerd (Evers et al., 2013). Enerzijds dient een onderscheid gemaakt te worden tussen de verschillende soorten betrouwbaarheid (interne consistentie, test-hertest betrouwbaarheid, interbeoordelaarsbetrouwbaarheid). Daarnaast is ook het verschil tussen tests bedoeld voor belangrijke en minder belangrijke beslissingen op individueel niveau van belang. Bij de beoordeling van de betrouwbaarheidscoëfficiënten worden voor het Totaal IQ en de indexscores de criteria

voor belangrijke beslissingen op individueel niveau toegepast. Voor subtestscores gelden de criteria voor minder belangrijke beslissingen op individueel niveau. Het is namelijk niet gerechtvaardigd om op basis van één subtest een uitspraak te doen over een index.

De **interne consistentie** werd bepaald aan de hand van (de gecorrigeerde) Guttman lambda2. De betrouwbaarheidscoëfficiënten van de primaire en secundaire indexen zijn voor alle leeftijdsgroepen ≥ .80 (voldoende), en vaak ≥ .90 (goed). Op subtestniveau varieerden de betrouwbaarheidscoëfficiënten van ≥ .70 (voldoende) tot ≥ .80 (goed), met uitzondering van de primaire subtests *Woordenschat* en *Begrijpen,* die bij respectievelijk één en twee leeftijdsgroepen betrouwbaarheidscoëficiënten hadden die < .70 (onvoldoende) waren.

Omtrent het bepalen van de interne consistentie worden nog volgende opmerkingen meegegeven.

- Aangezien de subtests *Symbool Substitutie Coderen*, *Symbool Zoeken* en *Figuur Zoeken* gescoord worden op tijd kan Guttman lambda2 niet berekend worden, en dient de score opgesplitst te worden. Voor de subtests *Symbool Substitutie Coderen* en *Symbool Zoeken* werd een split-half (eerste helft versus tweede helft) voorgesteld als alternatief, waarbij de auteurs er van uit gaan dat de helften gelijkwaardig zijn. Meer evidentie met betrekking tot de item-moeilijkheid en het oefen- of leereffect zijn nodig om deze stelling hard te maken. De test in 4 splitsen had in deze misschien een beter alternatief geweest.
- Voor de subtest *Figuur Zoeken* werd een test-hertest voorgesteld, en dit voor twee groepen. Er wordt in de Technische handleiding geen argumentatie gegeven over de keuze van (a) waarom twee groepen, (b) waarom 6-9 jaar en 10-16 jaar. Waar de betrouwbaarheid voor de groep 9 jaar op .74 geschat wordt, ligt deze voor de groep 10 jaar op .83; wat toch een opvallend verschil is. Desgevraagd heeft de testauteur aangegeven dat dit pragmatische keuzes zijn geweest. Deze keuzes hebben echter ook een impact op de betrouwbaarheid voor de totale steekproef. Zo valt de betrouwbaarheid voor de totale steekproef nog net binnen de grenzen van 'goed'. De vraag kan gesteld worden in hoeverre een verdeling in andere leeftijdscategorieën de betrouwbaarheid beïnvloedt en in hoeverre de huidige verdeling het meest gunstige resultaat in termen van betrouwbaarheidscoëfficiënten oplevert.

Bij de **test-hertest betrouwbaarheid** werden 36 Nederlandse en 45 Vlaamse kinderen een tweede keer getest met de WISC-V-NL, met een gemiddeld tijdsinterval van twee maanden. Het onderzoek wees uit dat de stabiliteit van het Totaal IQ uitstekend was (≥ .90). Bij de indexscores varieerden de hertestbetrouwbaarheidscoëfficiënten tussen voldoende (≥ .70) en goed (≥ .80). Ook de coëfficiënten op het niveau van de subtests lagen in het algemeen tussen voldoende (≥ .60) en goed (≥ .70), met twee subtests met betrouwbaarheidscoëfficiënten die als uitstekend (≥ .80) beoordeeld kunnen worden. Bij de interpretatie van significante verschillen tussen subtestscores dienen de absolute waarden van de test-hertest coëfficiënten (i.e., < .80, met uitzondering van twee subtests) in het achterhoofd gehouden te worden. Moesten in de formules van de verschilsignificanties tussen alle subtests de hertestbetrouwbaarheden gehanteerd worden, zouden er in de praktijk nog maar weinig significante verschillen optreden. Bij het testen van verschillen tussen subtests wordt dan ook beter het .01-significantieniveau gehanteerd.

Uit de resultaten van het test-hertest betrouwbaarheidsonderzoek bleek verder dat er op bijna elke subtest (met uitzondering van twee) een leereffect voorkwam. Op basis van deze resultaten omtrent het leereffect wordt daarom aangeraden een tussentijd van minimaal één tot twee jaar aan te houden. Bij een eventuele hertesting binnen die periode hanteert men dus best een andere intelligentietest zoals de RAKIT-II, CoVaT-CHC (9j.6m. – 13j.11m.) of SON-R 6-40j. (alleen Gf en Gv).

Er kunnen verschillende tekorten opgemerkt worden over het uitgevoerde test-hertestbetrouwbaarheidsonderzoek. De test-herteststeekproef is klein. Dergelijk kleine steekproefgrootte wordt in de literatuur als onvoldoende beoordeeld (Evers et al., 2013). Verder geven de demografische gegevens (Tabellen 5.4 en 5.5) weer dat hoewel de steekproef min of meer gestratificeerd is volgens de populatiestreefpercentages, er tevens grote verschillen merkbaar zijn (e.g., streefpercentage vrouw 50%, steekproefpercentage vrouw 61.7%). Ook is het op basis van de huidige gegevens onmogelijk te weten of alle leeftijdsgroepen vertegenwoordigd waren in het onderzoek. Als laatste is de grote range van tijdsintervallen tussen beide testings opvallend (15-161 dagen). In de COTAN-richtlijnen wordt inzake het test-hertestinterval gesteld dat 'een zeer kort interval (tot enkele weken) in het algemeen niet zinvol is, vanwege herinneringseffecten. Bij een test die is bedoeld voor voorspelling op lange termijn is het zinvol een relatief lang test-hertestinterval te kiezen' (Evers et al., 2010, p. 36). Een test zoals de WISC-V valt eerder onder de instrumenten die voorspellingen op langere termijn beogen. De ondergrens van 15 dagen als minimale tussentijd lijkt op basis van bovenstaande redenen dan ook niet verantwoord.

De **interbeoordelaarsbetrouwbaarheid** werd eveneens getoetst. De subtests *Overeenkomsten*, *Woordenschat* en *Begrijpen* werden door drie beoordelaars onafhankelijk van elkaar en zonder dat ze de kinderen gezien hadden, gescoord. Het ging om 51 willekeurig gekozen protocollen waarvan de antwoorden volledig uitgeschreven waren. De resultaten lagen tussen .98 en .99. Dit is uitstekend. Wel dient opgemerkt te worden dat enkel de *scoring* van bepaalde subtests in termen van interbeoordelaarsbetrouwbaarheid werd onderzocht. De betrouwbaarheid van de *afname* kan niet onderzocht worden. Toch moet men zich er van bewust zijn dat zowel bij afname als scoring verschillen tussen testleiders kunnen optreden. Dit wordt bijvoorbeeld geïllustreerd door de interpretatiefout van de Vlaamse testleiders bij de subtests *Symbool Zoeken* en *Symbool Substitutie Coderen*. Het is daarom niet onbelangrijk om voldoende in te zetten op de training van de proefleiders opdat de testafname op een gestandaardiseerde manier kan gebeuren.

Aandachtspunten

- Het is aangewezen om de verschillen tussen subtesten te toetsen op het .01-significantieniveau.
- Verder worden volgende aandachtspunten, die eveneens vermeld worden in de Technische handleiding, benadrukt.
    - Hertesting met WISC-V-NL binnen 1 à 2 jaar wordt afgeraden. Men doet best beroep op een alternatieve intelligentietest.
    - Het hanteren van individuele subtestscores voor het nemen van belangrijke beslissingen op individueel niveau is niet gerechtvaardigd.
    - Training is van primordiaal belang om de betrouwbaarheid en standaardisatie van testafnames te waarborgen.

Wat de beoordeling van betrouwbaarheid betreft, werd de evaluatie opgesplitst voor de verschillende besproken aspecten (interne consistentie, test-hertestbetrouwbaarheid en interbeoordelaarsbetrouwbaarheid). De **interne consistentie en interbeoordelaars-betrouwbaarheid** worden als **goed** beoordeeld, de **test-hertestbetrouwbaarheid** als **onvoldoende**. Bovenstaande aandachtspunten zijn van toepassing.

Begripsvaliditeit

De Technische handleiding rapporteert over de inhoudsvaliditeit, responsprocessen, de interne structuur met analyse van de intercorrelaties, confirmatorische factoranalyse, relaties met andere tests, relaties met biografische gegevens, onderzoeken bij klinische en specifieke groepen en tenslotte de (gelijktijdige) criteriumvaliditeit.

Inhoudsvaliditeit en responsprocessen

De inhoudsvaliditeit en het onderzoek naar responsprocessen worden positief beoordeeld. Wel is het jammer dat er nergens gegevens beschikbaar zijn over de moeilijkheidsgraden van de items. Deze zijn nochtans uiterst belangrijk voor het juist functioneren van de stop- of teruggaanregels (cf. opmerking bij onderdeel *Kwaliteit van het testmateriaal*).

Interne structuur

De interne structuur wordt in de Technische handleiding theoretisch beargumenteerd. Op basis van de hedendaagse intelligentiestructuur wordt een vijffactoroplossing voorgesteld en wordt bijgevolg het klassieke onderscheid tussen Verbaal en Performaal IQ verlaten. De bespreking van de structuur blijft eerder beperkt, met enkel een uiteenzetting van de relaties tussen items, subtests en indexscores.

Om het aantal factoren *exploratief* te bepalen zijn nochtans enkele eenvoudige methodes mogelijk. Traditioneel wordt bijvoorbeeld gekeken naar de eigenwaarden van de variantie-covariantiematrix of correlatiematrix. Het aantal eigenwaarden groter dan 1 (Kaiser, 1960) wordt gezien als het aantal factoren/componenten dat kan geëxtraheerd worden. Een minder conservatieve grens is die van Jolliffe (1972), waarbij het aantal eigenwaarden groter dan .70 geteld wordt. Visueel wordt voorgesteld om de scree-plot te inspecteren: het aantal factoren komt overeen met waar de 'knie' zit (Cattell, 1966). Een andere, meer verfijnde techniek, is via parallel analyse, waarbij de structuur zoals bekomen in de sample wordt vergeleken met random samples. Tenslotte kan ook worden gekeken naar de cumulatieve proportie verklaarde variantie van de beschouwde componenten, waarbij een grens van 80% als goed kan worden beschouwd, 70% als voldoende (zie ook Revelle, 2017). Wanneer door ons op basis van de gerapporteerde correlatiematrix (Tabel 6.1) deze analyses uitgevoerd worden, wordt er weinig evidentie gevonden dat een vijffactoroplossing ideaal is. Enkel op basis van het 70% criterium kan een vijffactoroplossing verdedigd worden als zijnde goed. De argumentatie lijkt dus voornamelijk theoretisch van aard.

Intercorrelaties

Zoals beschreven in de Technische handleiding (p. 89) werden door de testauteur verschillende a-priori hypotheses opgesteld omtrent intercorrelaties tussen subtest-, index- en processcores. De gevonden resultaten bevestigen deze hypotheses, zij het niet steeds op een overtuigende manier. Wat vooral opvalt, is dat de subtests van de Verwerkingssnelheid Index globaal een zwakke correlatie met de andere subtests vertonen én er ook een eerder zwakke correlatie is van *Figuur Zoeken* met alle subtests, ook met andere subtests van de Verwerkingssnelheid Index. *Figuur Zoeken* lijkt daarom geen valabel alternatief voor de subtests *Symbool Substitutie Coderen* en *Symbool Zoeken*.

Op basis van de intercorrelatieanalyses en gezien de opgenomen testen voor de indices en het Totaal IQ kan de vooropgestelde structuur als voldoende worden beschouwd.

Confirmatorische factoranalyse

Uit de confirmatorische factoranalyse (CFA) blijkt dat de vooropgestelde modellen de data in het algemeen goed lijken te fitten. Gelet op de fit-indices van de CFA's kan de voorgestelde vijffactorstructuur als voldoende beschouwd worden. De vooropgestelde modellen vormen dus een goede representatie van de data, maar zijn geen verklaring, zoals beweerd wordt door de testauteur.

Verder kan opgemerkt worden dat de keuze van aparte leeftijdsgroepen beter beargumenteerd had kunnen worden. Aangezien meer (Nederlands-Vlaamse) data voorhanden is in de uiterste groepen, had de laatste groep 15-16 jaar kunnen zijn, in plaats van 14-16 jaar. De onderverdeling in 6-7 jaar, 8-9 jaar, 10-11 jaar, 12-14 jaar en 15-16 jaar lijkt een betere keuze geweest te zijn.

Relaties met andere testen

Het is jammer dat bij het onderzoek met andere tests alleen Wechsler varianten zijn gebruikt (WPPSI-III-NL, WISC-III-NL, WAIS-IV-NL). Gezien de grote interne gelijkenis lijkt dit onderzoek meer op een hertestingsbetrouwbaarheid dan een echt validiteitsonderzoek. Het was mooi geweest moest er onderzoek plaats gevonden hebben met bijvoorbeeld de RAKIT-II of SON-R 6-40.

Daarnaast kunnen nog enkele zaken opgemerkt worden omtrent de gehanteerde methodologie. Ten eerste wordt er in de handleiding gesteld dat de tests 'zo veel mogelijk' in counterbalanced order afgenomen zijn. Dit blijkt echter bij de WPPSI-III-NL niet succesvol geweest te zijn, aangezien de auteurs de contrabalancering (62-38%, in plaats van beoogde 50/50%) aanhalen als argument om te verklaren waarom het Totaal IQ van de WPPSI-III-NL hoger ligt dan het Totaal IQ van de WISC-V-NL (zie *Errata Technische handleiding* waarom dit argument een foutieve redenering is). De contrabalancering wordt verder niet besproken bij het onderzoek met de WISC-III-NL en de WAIS-IV-NL. Het is dus onduidelijk wat 'zoveel mogelijk' bij deze onderzoeken betekent en in hoeverre afwijkingen in deze contrabalancering de resultaten heeft beïnvloed. Ten tweede varieerde het tijdsinterval tussen beide afnames, zowel bij het onderzoek naar de relatie tussen de WISC-V-NL met de WISC-III-NL, als bij het onderzoek naar de relatie met de WPPSI-III-NL van 2 dagen tot 44 en 63 dagen respectievelijk. Gezien de inhoudelijke overlap tussen de tests kan de vraag gesteld worden of een tijdsinterval van 2 dagen niet te kort is om leereffecten uit te sluiten. Ter vergelijking: zowel bij het onderzoek met hoogbegaafde kinderen als kinderen met een verstandelijk beperking werd een grens van 6 maanden gehanteerd. De vroegere IQ-testing moest minstens 6 maanden geleden gebeurd zijn.

Relaties met biografische gegevens

Er wordt in de Technische handleiding ingegaan op de samenhang van een WISC-V-NL score met het opleidingsniveau van de moeder enerzijds en met het opleidingsniveau van het kind anderzijds. Deze relaties liggen in de lijn met wat verwacht wordt. De analyses blijven echter beperkt tot het beschrijven van de scoreverschillen, zonder dat uit de tekst valt af te leiden of deze verschillen verder statistisch getoetst zijn geweest. Het is dan ook onduidelijk welke van de geobserveerde verschillen significant zijn. De tekst blijft hier ambigu over. Er wordt bijvoorbeeld op p. 119 gesproken over een trend die de verwachtingen ondersteunt. Kunnen we hier dan uit concluderen dat dit verschil niet of marginaal significant is?

Onderzoeken bij klinische en specifieke groepen

Een bespreking van deze onderzoeken werd onder criteriumvaliditeit geplaatst (zie infra).

Aandachtspunten

- Er bestaat twijfel over het gebruik van de secundaire subtest *Figuur Zoeken* als vervanging voor de primaire subtest *Symbool Substitutie Coderen* bij het bepalen van het Totaal IQ.

De **begripsvaliditeit** werd als **voldoende** beoordeeld. Bovenstaand aandachtspunt is van toepassing.

Criteriumvaliditeit

De (gelijktijdige) criteriumvaliditeit werd nagegaan door te onderzoeken hoe goed de score op de WISC-V-NL kan discrimineren enerzijds tussen een groep kinderen met een verstandelijke beperking (N = 58) en een gematchte controlegroep en anderzijds tussen een groep hoogbegaafde kinderen (N = 27) en een gematchte controlegroep. Met een *area under the curve* (AUC) van 0.96% voor de groep met verstandelijke beperking en een AUC van 0.83% voor de hoogbegaafden heeft men een uitstekend en goed resultaat. Toch kunnen enkele opmerkingen gegeven worden over deze analyse, en meer specifiek over het onderzoek bij de groep hoogbegaafden:

- Ten eerste dient opgemerkt te worden dat een deel van de hoogbegaafde kinderen gediagnosticeerd is geweest aan de hand van een vorige versie van de WISC-V-NL (namelijk de WISC-III-NL) of een Wechsler leeftijdsvariant (WPPSI-R, WPPSI-III-NL). In hoeverre is er geen sprake van een confound bij deze analyse? De huidige WISC-V-NL score van een kind – die op basis van een vroegere WISC-III-NL score als hoogbegaafd is gediagnosticeerd – wordt als bewijs aangehaald om aan te tonen dat de WISC-V-NL goed differentieert tussen normaal en hoogbegaafde kinderen. Men kan in dit geval niet stellen dat de beoordeling van hoogbegaafdheid helemaal onafhankelijk van de WISC-V-NL gesteld is geweest.
- Ten tweede haalde de steekproef een gemiddeld Totaal IQ van 122.6, ondanks dat er een selectie gemaakt werd op een IQ van minstens 130 (gemiddeld 139). Dit is ruim één SD lager. Regressie naar het gemiddelde wordt als mogelijke verklaring aangehaald, alsook de aanwezigheid van vier outliers, met een IQ onder 115. Cijfers zonder deze outliers worden niet gerapporteerd. De mediaan was in deze ook een betere statistiek geweest.
- Ook is de steekproefgrootte van het onderzoek met hoogbegaafde kinderen onvoldoende (Evers et al., 2013).
- Als laatste kan in aansluiting met wat in dit rapport aan bod kwam bij *Normen. Virtueel karakter extreem hoge en lage IQ-scores* (zie de 1.2% te weinig) de vraag gesteld worden of het virtuele karakter aan het einde van de IQ-schaal een mogelijk knelpunt camoufleert. Dit lijkt toch een aandachtspunt voor verder onderzoek bij deze betrokken doelgroepen. De vraag kan gesteld worden of bij de onderzochte klinische groepen daadwerkelijk IQ-scores < 50 en > 150 voorkwamen. Dit kan niet opgemaakt worden uit de gegevens in de tabellen 6.9 en 6.10.

De hierboven beschreven analyse biedt ook enkel evidentie voor gelijktijdige criteriumvaliditeit. Zoals in de handleiding aangegeven wordt, is meer onderzoek nodig waarbij ook de voorspellende waarde van de WISC-V wordt nagegaan.

De Technische handleiding stelt verder: "*aanvullend bewijs voor de criteriumvaliditeit wordt verkregen uit onderzoek naar bijzondere doelgroepen*" (p. 122). Dit is terecht, maar het lijkt ons dat niet elke

bijzondere doelgroep daarvoor in aanmerking komt. Dat de doelgroepen hoogbegaafden en kinderen met een lichte tot matige verstandelijke beperking aan bod komen, is evident. Kinderen met ADHD of een autismespectrumstoornis hoeven echter per definitie geen criteriumgroepen te zijn. De meerwaarde van deze onderzoeken kan dan ook in vraag gesteld worden. Er is sprake van comorbiditeit binnen de Nederlands-Vlaamse klinische groepen en er is een grote mate van variabiliteit binnen de klinische diagnoses in termen van verschijningsvorm en symptomatologie. In hoeverre zijn deze resultaten dan betekenisvol? Het geeft de foutieve indruk dat de gevonden groepseffecten generaliseerbaar zijn naar individuele cliënten, en dat een prestatie op de WISC-V-NL een diagnose ADHD of ASS al dan niet kan ondersteunen. Wij willen dan ook vragen dat bij paragraaf 6.7.3 en 6.7.4 een waarschuwing kan toegevoegd worden zoals ook op p. 122 §3 gebeurd is.

Aandachtspunten

- De criteriumvaliditeit dient nog verder onderzocht te worden.
- De in de Technische handleiding gepresenteerde resultaten van klinische groepen (ASS, ADHD) zijn niet geheel representatief. De onderzoeken zijn exemplarisch bedoeld en dienen geen klinisch doel. De gevonden groepsverschillen op IQ mogen nooit gehanteerd worden om bij individuele cliënten een diagnose of classificatie te stellen.

Op basis van bovenstaande opmerkingen werd de criteriumvaliditeit als **onvoldoende** beoordeeld. Bovenstaande aandachtspunten dienen in rekening te worden gebracht.

Samenvatting beoordeling

Het voorliggende reviewrapport heeft betrekking op de evaluatie van de WISC-V-NL, en meer specifiek op de Vlaamse normeringsdata inzake. Daarnaast werd de WISC-V-NL ook beoordeeld door de COTAN. Beide evaluaties gebeurden onafhankelijk van elkaar. Het rapport van de COTAN kan bekomen worden via: https://www.cotandocumentatie.nl/beoordelingen/b/15651/wechsler-intelligence-scale-for-children-fifth-edition/. Na consultatie van dit oordeel blijkt dat beide evaluaties gelijklopend zijn. Het is aan te bevelen ook het COTAN-rapport te consulteren om een genuanceerd beeld te krijgen van de kwaliteit van de gecombineerde Nederlands-Vlaamse normen.

Onderstaande kwaliteitslabels werden door het Kwaliteitscentrum voor Diagnostiek vzw toegekend bij de beoordeling van het gebruik van de WISC-V-NL in Vlaanderen. Bij elk onderdeel werden in het reviewrapport tevens aandachtspunten geformuleerd. Er wordt geadviseerd deze aandachtspunten te consulteren om te komen tot een verantwoord gebruik van de WISC-V-NL in Vlaanderen.

Beoordeling WISC-V-NL voor gebruik in Vlaanderen

Uitgangspunten testconstructie	Goed
Kwaliteit van het testmateriaal	Goed
Kwaliteit van de handleiding	Goed
Normen	Voldoende
Betrouwbaarheid	
Interne consistentie	Goed
Test-hertestbetrouwbaarheid	Onvoldoende
Interbeoordelaarsbetrouwbaarheid	Goed
Begripsvaliditeit	Voldoende
Criteriumvaliditeit	Onvoldoende

Errata Technische handleiding

- p. 78: De gehanteerde definitie van betrouwbaarheidsintervallen is niet correct. De betekenis is dat bij herhaling van de procedure, met steeds nieuwe (aselecte) steekproeven uit dezelfde populatie, verwacht mag worden dat bijvoorbeeld 90% of 95% van de zo berekende intervallen de parameter μ zullen bevatten.
- p. 79: laatste paragraaf: st<u>r</u>eekproef
- p. 87: “Als gevolg hiervan werden er vier indexscores voor de Amerikaanse WISC-III en vier indexscores voor de WAIS-III geconstateerd die specifieke cognitieve vaardigheden vertegenwoordigden: Verbaal Begrip, Perceptuele Organisatie, Vrijheid van Afleidbaarheid (of Werkgeheugen voor de WAIS-III) en ~~Verbaal Begrip~~, naast de traditionele VIQ-, PIQ- en TIQ-scores.”
- p. 92: “De resultaten van exploratieve factoranalyse kunnen namelijk vrij gevoelig ~~kunnen~~ zijn voor kleine verschillen in het patroon van correlaties.”
- P. 102-103: Het gemiddeld Totaal IQ van de WPPSI-III-NL ligt meer dan 4 punten hoger dan het Totaal IQ van de WISC-V-NL. Als mogelijke verklaring wordt hiervoor het leereffect aangehaald. Echter, bij 62% van de kinderen werd eerst de WPPSI-III-NL afgenomen, bij 38% eerst de WISC-V-NL. Indien er sprake is van een leereffect, lijkt een hogere score voor de WISC-V-NL eerder waarschijnlijk, aangezien voor 68% deze test later kwam.
- p. 153: “Een discrepantie waarbij VsI <u>~~kleiner~~ groter</u> is dan WgI wijst erop dat het kind mogelijk beter is in het nemen van snelle beslissingen met informatie die is opgeslagen in het kortetermijngeheugen dan in het manipuleren van die informatie”.

Referenties

Bechger, T., Hemker, B., & Maris, G. (2009). *Over het gebruik van continue normering.* Arnhem: Cito.

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245–276.

Evers, A., Hagemeister, C., Hostmaelingen, A., Lindley, P., Muniz, J., & Sjöberg, A. (2013). *EFPA review model for the description and evaluation of psychological and educational tests. Test review form and notes for reviewers version 4.2.6.* Geraadpleegd op 12 september 2018 via http://www.efpa.eu/professional-development/assessment.

Evers, A., Lucassen, W., Meijer, R.R., & Sijtsma, K. (2010). *COTAN Beoordelingssysteem voor de kwaliteit van tests.* Amsterdam: NIP.

Evers, A.V.A.M., Sijtsma, K., Lucassen, W., & Meijer, R.R. (2010). The Dutch review process for evaluating the quality of psychological tests: History, procedure and results. *International Journal of Testing, 10*(4), 295-317.

Flanagan, D.P., & Alfonso, V.C. (2017). *Essentials of WISC-V Assessment.* Hoboken, New-Jersey-USA: J. Wiley & Sons, Inc.

Hendriks, M.P.H., Ruiter, S., Schittekatte, M., & Bos, A. (2018). *Wechsler Intelligence Scale for Children – Fifth edition – Nederlandstalige bewerking. Technische handleiding.* Amsterdam: Pearson Benelux B.V.

Jolliffe, I.T. (1972). Discarding variables in a principal component analysis. I: Artificial data. *Applied Statistics*, *21*, 160-173.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement*, *20*(1), 141-151.

Kaufman, A.S., Riaford, S.E., & Coalson, D.L. (2016). *Intelligent Testing with the WISC-V.* Hoboken, New-Jersey-USA: J. Wiley & Sons, Inc.

Revelle, W.R. (2017). psych: Procedures for Personality and Psychological Research. Software.

Van Baar, A.L., Steenis, L.J.P., Verhoeven, M., Hessen, D.J., & Smits-Engelsman, B.C.M. (2015). *Bayley-III-NL. Technische handleiding, tweede herziene druk.* Amsterdam: Pearson Assessment and Information B.V.

Weis, L.G., Saklofske, D.H., Holdnack, J.A., & Prifitera, A. (2016). *WISC-V Assessment and Interpretation Scientist-Practitioner Perspectives.* London: Academic Press, Elsevier.